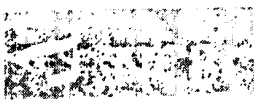

National Library of Medicine

PubMed

1: *Methods Enzymol* 1996;266:295-322

Protein sequence comparison at genome scale.

Koonin EV, Tatusov RL, Rudd KE

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, U.S.A.

An adequate set of computer procedures tailored to address the task of genome-scale analysis of protein sequences will greatly increase the beneficial impact of the genome sequencing projects on the progress of biological research. This is especially pertinent given the fact that, for model organisms, one-half or more of the putative gene products have not been functionally characterized. Here we described several programs that may comprise the core of such a set and their application to the analysis of about 3000 proteins comprising 75% of the E. coli gene products. We find that the protein sequences encoded in this model genome are a rich source of information, with biologically relevant similarities detected for more than 80% of them. In the majority of cases, these similarities become evident directly from the results of BLAST searches. However, methods for motif analysis provide for a significant increase in search sensitivity and are particularly important for the detection of ancient conserved regions. As a result of sequence similarity analysis, generalized functional predictions can be made for the majority of uncharacterized ORF products, allowing efficient focusing of experimental effort. Clustering of the E. coli proteins on the basis of sequence similarity shows that almost one-half of the bacterial proteins have at least one paralog and that the likelihood that a protein belongs to a small or a large cluster depends on the function of this particular protein.

PMID: 8743691